Ad hoc compounds in English to Finnish machine translation

Hurskainen, Arvi

University of Helsinki, Institute for Asian and African Studies

2020

Hurskainen , A 2020 ' Ad hoc compounds in English to Finnish machine translation ' Technical Reports on Language Technology , no. 57 , University of Helsinki, Institute for Asian and African Studies , Helsinki . < http://www.njas.helsinki.fi/salama/ad-hoc-compounds-in-en-to-fi-mt.pdf >

http://hdl.handle.net/10138/317504

Technical Reports on Language Technology
Report No 57, 2020
http://www.njas.helsinki.fi/salama

Ad hoc compounds in English to Finnish machine translation

Arvi Hurskainen
Department of Languages, Box 59
FIN-00014 University of Helsinki, Finland
arvi.hurskainen@helsinki.fi

Abstract

Language is in a continuous development process, which poses a challenge to machine translation. People also tend to have their only styles of writing, which is often difficult to translate. A special problem is the so called *ad hoc* compounding, where the writer decides to construct such compounds, which do not exist in any dictionary. By compounds I mean here such clusters of words, which are written together, using '-' as a connecting diacritic, not the normal compounding, where members of a compound are written as separate words. In addition to ad hoc compounds, there are also established compounds using the same connecting method. This report describes the methods for handling these compounds in English to Finnish machine translation.

Key Words: *compounding, morphological analysis, machine translation.*

1 Introduction

Current English has a common trend to construct such compounds, where two or more words are written together using dash as a connecting diacritic. In the absence of any better term, I call them ‘dash compounds’. Because of brevity, in this report I use the term ‘compound’ for meaning dash compounds, although in normal usage the term ‘compound’ means the cluster of separate English words.

The normal procedure in handling these compounds is to list them directly into the morphological lexicon along with other words. When the compounds are translated, each compound must also be separately listed into the morphological lexicon. If such compounds would form a closed list, they could be handled in this way, along with normal words. However, new compounds appear continually, and a sufficiently comprehensive list cannot be maintained. There are also compounds, where the first element is a varying number, such as *76-year*, *84-year-old*, *13-hour*, *24-metre*, and *45-minute*. Therefore, an additional method should be found.

One approach would be that we allow all nouns, adjectives and adverbs combine as single units. This method, however, would be too productive, and removing the wrong combinations would be uneconomical. Therefore, we make a list of such words, which appear as last elements of the compound. These words are either nouns or adjectives.

The idea in this approach is that we limit the combinations to such words, which we already know to appear as last part of compounds.

On this point we must consider the translation of the compound. When we translate into Finnish, we also must take into account the fact that any of the compound members

may inflect in Finnish. The management of the inflection of both members is not a simple matter, because the case of the first member depends on the last member of the compound, and the case of the last member of the compound depends on the sentence context.

The difference between the first and last part of the compound is that for the last part we can construct a closed list, but for the first part such a list is not feasible. The closed list, however, is not fully inclusive, but it includes the most often occurring cases. In designing the closed lists, we can indicate the case of the first part, whatever the first part is. It is in all contexts the same, because the case of the first part is defined by the last part. The case of the last part varies according to context, and the inflection tags for it are added on the basis of context.

The Finnish glosses for the last part of the compound can be added either by adding the glosses directly into the morphological lexicon, or in conjunction with the normal gloss adding after morphological analysis and disambiguation. The first method is easier to implement, and in this report I will describe that solution.

The compounds that we discuss here function usually as adjectives. This is the case regardless the POS class of each compound member. However, we discuss the compounds in two groups, depending on the POS category of the last member.

2 Ad hoc compounds with adjective as last member

As I said in the introduction, I have included the lists of the last part the compound as two separate sub-lexicons in the morphological lexicon. One of these sub-lexicons contains nouns, and the other one contains adjectives. In (1) is the list of adjectives.

(1)

```
-backed # "= <GEN [ tukema N10 ]";
-baked # "= [ leivottu N1-C ] ";
-based # "= <ILL [ pohjainen N38 <NOM , perustuva N10 ]";
-catching # "= <ILL [ pistävä N10 FRONT ]";
-causing # "= <PAR [ aiheuttava N10 ]";
-century # "= [ vuosisata N9-F ]";
-changing # "= <PAR [ muuttava N10 ]";
-clad # "= <NOM [ pukuinen N38 , päällysteinen N38 FRONT ]";
-class # "= <NOM [ luokkainen N38 ]";
-controlled # "= <GEN [ kontrolloima N10 ]";
-deep # "= <GEN [ syvyinen N38 FRONT ]";
-degree # "= <GEN [ asteen ]";
-dominated # "= <GEN [ hallitsema N10 ]";
-elect # "= [ valittu N1-C ]";
-faced # "= <NOM [ kasvoinen N10 ]";
-filled # "= <GEN [ täyttämä N10 FRONT ]";
-free # "= <NOM [ vapaa N17 ]";
-fuelled # "= <NOM [ käyttöinen N38 FRONT ]";
-funded # "= <GEN [ rahoittama N10 ]";
-game # "= <GEN [ pelinen N38 FRONT ]";
-handed # "= <NOM [ kätinen N38 FRONT ]";
-held # "= <GEN [ pitämä N10 FRONT ]";
```

```
-inspired # "= <GEN [ innoittama N10 ]";
-led # "= <NOM [ johtoinen N38 ]";
-level # "= <NOM [ tasoinen N38 ]";
-like # "=  <GEN [ tapainen N38 ]";
-lived # "= <NOM [ ikäinen N38 ]";
-living # "= <NOM [ ikäinen N38 ]";
-looking # "= <GEN [ näköinen N38 FRONT ]";
-long # "= <GEN [ pituinen N38 ]";
-made # "= <INE [ tekoinen N38 <NOM , tehty N1-F FRONT ]";
-man # "= <NOM [ miehinen N38 FRONT ]";
-minded # "= <NOM [ mielinen N38 ]";
-needed # "= [ tarvittu N1-C ]";
-year-old # "= [ vuotta vanha N9 ]";
-old # "= <PAR [ vanha N9 ]";
-olds # "= <PAR [ vanha N9 ] PL";
-operated # "= <NOM [ käyttöinen N38 FRONT ]";
-owned # "= <NOM [ omisteinen N38 ]";
-page # "= <NOM [ sivuinen N38 ]";
-paid # "= [ maksettu N1-C ]";
-person # "= [ henkinen N38 FRONT ]";
-point # "= [ kohtainen N38 ]";
-ranked # "= <ALL [ arvioitu N1-F ]";
-ranking # "= <ALL [ arvioiva N10 ]";
-registered # "= [ rekisteröity N1-F FRONT ]";
-related # "= <ILL [ liittyvä N10 FRONT ]";
-rich # "= <NOM [ rikas N41-A ]";
-round # "= [ kierroksinen N38 ]";
-scale # "= [ luokkainen N38 ]";
-shaped # "= <GEN [ muotoinen N38 ]";
-sided # "= [ sivuinen N38 ]";
-sized # "= <GEN [ kokoinen N38 ]";
-sourced # "= <NOM [ lähtöinen N38 FRONT ]";
-sponsored # "= <GEN [ sponsoroima N10 ]";
-stricken # "= <GEN [ lyömä N10 FRONT ]";
-strong # "= <GEN [ vahvuinen N38 ]";
-style # "= <NOM [ tyylinen N38 FRONT ]";
-tall # "= <GEN [ korkuinen N38 ]";
-term # "= [ ajan ]";
-thick # "= <GEN [ paksuinen N38 ]";
-time # "= [ kertainen N38 , aikainen N38 ]";
-traded # "= [ ostettu N1-C ]";
-wide # "= <GEN [ laajuinen N38 ]";
-winning # "= <PAR [ voittava N10 ]";
-won # "= [ voitettu N1-C ]";
-working # "= [ työskentelevä N10 FRONT ]";
```

The list in (1) is an extract from the morphological lexicon. On each line, the first string is the last part of the compound, and it can be suffixed to any noun, adjective, numeral, or adverb. The hatch '#' means that the word formation stops here. The equal sign '=' means that the leftmost word is copied as gloss. The tag <GEN and other similar tags

indicate the case of the first part, whatever it is. Between square brackets is the gloss and its inflection code. In case the word has a front vowel inflection, it has the tag FRONT.

When we perform the plain analysis, without any further processing, we get results as in (2).

(2)
```
"<state-backed>"
      "state-backed" PREFR A  <GEN [ tukema N10 ]
"<state-controlled>"
      "state-controlled" PREFR A  <GEN [ kontrolloima N10 ]
"<hey-filled>"
      "hey-filled" A  <GEN [ täyttämä N10 FRONT ]
"<waist-deep>"
      "waist-deep" A  <GEN [ syvyinen N38 FRONT ]
```

If we compare the readings with the list in (1), we see that if the compound is interpreted as adjective, it gets a gloss of the last part. The gloss of the first part is not yet there. We add the gloss of the first part (3).

(3)
```
"<state-backed>"
      "state { valtio N3 , valtio-- COMP , valtion-- COMP , tila
N9 }-backed" PREFR A <GEN [ tukema N10 ]
"<state-controlled>"
      "state { valtio N3 , valtio-- COMP , valtion-- COMP , tila
N9 }-controlled" PREFR A <GEN [ kontrolloima N10 ]
"<hey-filled>"
      "hey { heinä N10 FRONT }-filled" A <GEN [ täyttämä N10 FRONT
]
"<waist-deep>"
      "waist { vyötärö N2 FRONT , vyötärönympärys N39 FRONT }-
deep" A <GEN [ syvyinen N38 FRONT ]
```

When we added the glosses of the first part above, we notice that there is often more than one gloss. In the case of *valtio*, there are gloss alternatives for the kind of compound that we need here. The gloss alternatives with the tag COMP are there for translating normal compounds, where members are written as separate words. We could select the appropriate alternative, that is *valtion*-- and leave the others out. This is the genitive form as requires the tag <GEN.

Because such ready forms are only in some common nouns, it is better to select the first gloss and inflect it using the general rules. When we take the first gloss and remove the rest, the result is as in (4).

(4)
```
"<state-backed>"
      "state-backed" { valtio N3 } PREFR A <GEN [ tukema N10 ]
"<state-controlled>"
      "state-controlled" { valtio N3 } PREFR A <GEN [ kontrolloima
N10 ]
```

```
"<hey-filled>"
      "hey-filled" { heinä N10 } A <GEN [ täyttämä N10 FRONT ]
"<waist-deep>"
      "waist-deep" { vyötärö N2 } A <GEN [ syvyinen N38 FRONT ]
```

The tag <GEN in each reading means that the first part must be in genitive. We first mark each gloss with ':' to show the point, where inflection starts (5).

(5)

```
"<state-backed>"
      "state-backed" { valti:o N3 } PREFR A <GEN [ tukem:a N10 ]
"<state-controlled>"
      "state-controlled"  {  valti:o  N3  }  PREFR  A  <GEN  [
kontrolloim:a N10 ]
"<hey-filled>"
      "hey-filled" { hein:ä N10 } A <GEN [ täyttäm:ä N10 FRONT ]
"<waist-deep>"
      "waist-deep" { vyötärö: N2 } A <GEN [ syvyi:nen N38 FRONT ]
```

Then we add the genitive suffix to the stem (6). Note that the original ending is still there, and it will be removed later.

(6)

```
"<state-backed>"
      "state-backed" { valti:o+on :N3 } PREFR A [ tukem:a N10 ]
"<state-controlled>"
      "state-controlled"  {  valti:o+on  :N3  }  PREFR  A  [
kontrolloim:a N10 ]
"<hey-filled>"
      "hey-filled" { hein:ä+än :N10 } A [ täyttäm:ä N10 FRONT ]
"<waist-deep>"
      "waist-deep" { vyötärö:+n :N2 } A [ syvyi:nen N38 FRONT ]
```

The noun alternative does not get inflection, because there is no tag for it. After pruning the result, we get the translation (7).

(7)
valtion tukema
valtion kontrolloima
heinän täyttämä
vyötärön syvyinen

When we put the compound into context, we see that also the last part inflects (8).

(8)

```
"<He>"
      "he" { hän Np9 FRONT } %SUBJ OUT HUM MALE CAPINIT PRON PERS
SG3 NOM
"<works>"
```

```
      "work" { toimia V61 } %+FMAINV V-1INF-TRA V PRES  SG
"<in>"
      "in" { M-LOC1 } %ADVL PREP
"<state-controlled>"
      "state-controlled" { valtion :N3 } PREFR A [ kontrolloim:a
N10 ] INE
"<company>"
      "company" { yritys N39 FRONT } %<P INDEF N SG INE
"<.>"
      "." { . }
```

We see that the system has added the tag INE on two last words. The tags are converted to surface form using the inflection code of each word (9).

(9)

```
"<He>"
      "he" { hän Np9 FRONT } %SUBJ OUT HUM MALE CAPINIT PRON PERS
SG3 NOM
"<works>"
      "work" { toimii V61 } %+FMAINV V-1INF-TRA V PRES  SG
"<in>"
      "in" { M-LOC1 } %ADVL PREP
"<state-controlled>"
      "state-controlled" { valtion :N3 } PREFR A [ kontrolloimassa
N10 ] INE
"<company>"
      "company" { yrityksessä N39 FRONT } %<P INDEF N SG INE
"<.>"
      "." { . }
```

The translation is in (10).

(10)
Hän toimii valtion kontrolloimassa yrityksessä.

3 Ad hoc compounds with noun as last member

In the second group of compounds, the last member is a noun. Yet the compound usually has an adjectival meaning. Nouns likely to be in the position of the last member of the compound are listed in (11).

(11)

```
-day # "= [ päiväinen N38 FRONT , <GEN päivän ]";
-floor # "= [ kerroksinen N38 , <GEN kerroksen ]";
-foot # "= [ jalkainen N38 , <GEN jalan ]";
-goal # "= [ päämääräinen N38 FRONT , <GEN päämäärän ]";
-hour # "= [ tuntinen N38 , <GEN tunnin ]";
-inch # "= [ tuumainen N38 , <GEN tuuman ]";
-metre # "= [ metrinen N38 FRONT , <GEN metrin ]";
```

```
-million # "= [ miljoonainen N38 , <GEN miljoonan ]";
-minute # "= [ minuuttinen N38 , <GEN minuutin ]";
-month # "= [ kuukautinen N38 , <GEN kuukauden ]";
-page # "= [ sivuinen N38 , <GEN sivun ]";
-pound # "= [ <GEN punnan ]";
-stage # "= [ vaiheinen N38 , <GEN vaiheen ]";
-step # "= [ askeleinen N38 , <GEN askeleen ]";
-store # "= [ kerroksinen N38 , <GEN kerroksen ]";
-storey # "= [ kerroksinen N38 , <GEN kerroksen ]";
-story # "= [ kerroksinen N38 , <GEN kerroksen ]";
-week # "= [ viikkoinen N38 , <GEN viikon ]";
-year # "= [ vuotinen N38 , <GEN vuoden ]";
-years # "year [ vuotinen N38 , <GEN vuoden ] PL";
```

We see that most of the words in the list express some measure. The words of the list are located as a separate sub-lexicons in the morphological analyser, and they can be suffixed to nouns, adjectives, numerals and adverbs.

The word has usually two glosses. The first gloss is an adjective and it does not require the inflection of the first member. The compound itself inflects according to context. Also, the two parts of the compound are usually written together as a single word.

The second gloss is a genitive form of the word, and it has no other forms. It requires that also the first part is in genitive.

Which of the two alternatives should be used in each case is not quite clear. Some allow both alternatives, but most often only one of them is appropriate. The selection requires context-sensitive disambiguation.

A set of examples in (11) shows how the analysis result looks like.

(11)

```
"<32-year>"
      "32-year" A  [ vuotinen N38 , <GEN vuoden ]

"<six-year>"
      "six-year" A  [ vuotinen N38 , <GEN vuoden ]

"<8-week>"
      "8-week" A  [ viikkoinen N38 , <GEN viikon ]

"<12-hour>"
      "12-hour" A  [ tuntinen N38 , <GEN tunnin ]

"<eight-week>"
      "eight-week" A  [ viikkoinen N38 , <GEN viikon ]
```

For adjective alternatives, the glosses were inserted, because they were listed in the sub-lexicon. The noun alternatives are without gloss.

We add glosses to the rest of words (12).

(12)
```
"<32-year>"
      "32-year" A [ vuotinen N38 ]
      "32-year" A <GEN [ vuoden ]
"<six-year>"
      "six { kuusi N27 , -kuusi N27 COMB , -kuudes N45 COMB ORD }
NUM-PL-year" A [ vuotinen N38 ]
      "six { kuusi N27 , -kuusi N27 COMB , -kuudes N45 COMB ORD }
NUM-PL-year" A <GEN [ vuoden ]
"<8-week>"
      "8-week" A [ viikkoinen N38 ]
      "8-week" A <GEN [ viikon ]
"<12-hour>"
      "12-hour" A [ tuntinen N38 ]
      "12-hour" A <GEN [ tunnin ]
"<eight-week>"
      "eight { kahdeksan N10b , -kahdeksan N10b COMB , -kahdeksas
N45 COMB ORD } NUM-PL-week" A [ viikkoinen N38 ]
      "eight { kahdeksan N10b , -kahdeksan N10b COMB , -kahdeksas
N45 COMB ORD } NUM-PL-week" A <GEN [ viikon ]
```

We see that if the first member is a number, no gloss is added. If it is a numeral, the corresponding gloss is added. Also, the two glosses with relevant tags were moved to their own lines.

We only need the first gloss for our purposes, and the rest can be removed (13).

(13)
```
"<32-year>"
      "32-year" A [ vuotinen N38 ]
      "32-year" A <GEN [ vuoden ]
"<six-year>"
      "six-year" { kuusi N27 } NUM-PL A [ vuotinen N38 ]
      "six-year" { kuusi N27 } NUM-PL A <GEN [ vuoden ]
"<8-week>"
      "8-week" A [ viikkoinen N38 ]
      "8-week" A <GEN [ viikon ]
"<12-hour>"
      "12-hour" A [ tuntinen N38 ]
      "12-hour" A <GEN [ tunnin ]
"<eight-week>"
      "eight-week" { kahdeksan N10b } NUM-PL A [ viikkoinen N38 ]
      "eight-week" { kahdeksan N10b } NUM-PL A <GEN [ viikon ]
```

Now all members except for the numbers have glosses. We copy the number from the source text to the target text (14).

(14)
```
"<32-year>"
      "32-year" A [ 32-vuotinen N38 ]
      "32-year" A <GEN [ 32-vuoden ]
```

```
"<six-year>"
      "six-year" { kuusi N27 } NUM-PL A [ vuotinen N38 ]
      "six-year" { kuusi N27 } NUM-PL A <GEN [ vuoden ]
"<8-week>"
      "8-week" A [ 8-viikkoinen N38 ]
      "8-week" A <GEN [ 8-viikon ]
"<12-hour>"
      "12-hour" A [ 12-tuntinen N38 ]
      "12-hour" A <GEN [ 12-tunnin ]
"<eight-week>"
      "eight-week" { kahdeksan N10b } NUM-PL A [ viikkoinen N38 ]
      "eight-week" { kahdeksan N10b } NUM-PL A <GEN [ viikon ]
```

The second reading in each cohort requires that the first member must be in genitive. This is implemented in (15).

(15)

```
"<32-year>"
      "32-year" A [ 32-vuoti:nen N38 ]
      "32-year" A <GEN [ 32-vuoden ]
"<six-year>"
      "six-year" { kuu:si :N27 } NUM-PL A [ vuoti:nen N38 ]
      "six-year" { kuu:si+den :N27 } NUM-PL A [ vuoden ]
"<8-week>"
      "8-week" A [ 8-viikkoi:nen N38 ]
      "8-week" A <GEN [ 8-viikon ]
"<12-hour>"
      "12-hour" A [ 12-tunti:nen N38 ]
      "12-hour" A <GEN [ 12-tunnin ]
"<eight-week>"
      "eight-week" { kahdeks:an :N10b } NUM-PL A [ viikkoi:nen N38
]
      "eight-week" { kahdeks:an+an :N10b } NUM-PL A [ viikon ]
```

When we proceed in producing the surface form, we must consider, whether or not the members will be written together in Finnish. The solution is different in the first and second cohort (16).

(16)

```
"<32-year>"
      "32-year" A [ 32-vuoti:nen N38 ]
      "32-year" A <GEN [ 32 vuoden ]
"<six-year>"
      "six-year" { kuu:sivuoti:nen N38 } A
      "six-year" { kuu:si+den :N27 } NUM-PL A [ vuoden ]
"<8-week>"
      "8-week" A [ 8-viikkoi:nen N38 ]
      "8-week" A <GEN [ 8 viikon ]
"<12-hour>"
      "12-hour" A [ 12-tunti:nen N38 ]
```

```
      "12-hour" A <GEN [ 12 tunnin ]
"<eight-week>"
      "eight-week" { kahdeks:anviikkoi:nen N38 } A
      "eight-week" { kahdeks:an+an :N10b } NUM-PL A [ viikon ]
```

In the first cohort, the members were written together. In the second cohort, the members were written as separate words. The POS tag A in the second cohort is wrong, because this is not an adjectival structure. It must be rewritten as noun (17).

(17)
```
"<32-year>"
      "32-year" A [ 32-vuoti:nen N38 ]
      "32-year" N <GEN [ 32 vuoden ]
"<six-year>"
      "six-year" { kuu:sivuoti:nen N38 } A
      "six-year" { kuu:si+den :N27 } NUM-PL N [ vuoden ]
"<8-week>"
      "8-week" A [ 8-viikkoi:nen N38 ]
      "8-week" N <GEN [ 8 viikon ]
"<12-hour>"
      "12-hour" A [ 12-tunti:nen N38 ]
      "12-hour" N <GEN [ 12 tunnin ]
"<eight-week>"
      "eight-week" { kahdeks:anviikkoi:nen N38 } A
      "eight-week" { kahdeks:an+an :N10b } NUM-PL N [ viikon ]
```

The translation of these structures without context is in (18).

(18)
32-vuotinen
32 vuoden

kuusivuotinen
kuuden vuoden

8-viikkoinen
8 viikon

12-tuntinen
12 tunnin

kahdeksanviikkoinen
kahdeksan viikon

We put some of these examples into context to see how they behave (19).

(19)
```
"<He>"
```

```
      "he" { hän Np9 FRONT } %SUBJ OUT HUM MALE CAPINIT PRON PERS
SG3 NOM
"<was>"
      "be" { olla V67b } %+FMAINV V-3INF-ILL O-LOC1 V PAST  SG
"<in>"
      "in" { M-LOC1 } %ADVL PREP
"<six-year>"
      "six-year" { kuusivuotinen N38 } A
"<war>"
      "war" { sota N10-F } %<P MIL DEF N  SG INE
"<.>"
      "." { . }
"<<s>>"
      "<s>" { <s> }
"<six-year>"
      "six-year" { kuuden :N27 } NUM-PL N [ vuoden ]
"<period>"
      "period" { ajanjakso N1 } %SUBJ OUT DEF N  SG NOM
"<has>"
      "have" { olla V67b } %+FAUXV HAVE-PERF V PRES  SG
"<ended>"
      "end" { päättyä V52-C FRONT } %-FMAINV V EN-PERF SG @SG
"<.>"
      "." { . }
```

In the first example, the adjective alternative *kuusivuotinen* was chosen. In the second example, the noun alternative *kuuden vuoden* was chosen.

We must note that it is difficult to find criteria for the correct choice, and sometimes both alternatives are correct.

There are also cases, where neither of these two alternatives is correct. Consider the examples in (20).

(20)
five-year plan - viisivuotinen suunnitelma
viiden vuoden suunnitelma
viisivuotissuunnitelma

The convention is that the last one is used. The example shows that it is not possible to handle all cases with only two gloss alternatives. If the third alternative would be added, the third reading could be selected in some cases. But would this be economical? The solution would require more rather unreliable rules. Perhaps the best solution is to list such difficult cases directly into the lexicon. I will discuss such problems in the last section.

4 Combination of two approaches

Above I have described the general approach for handling *ad hoc* compounds. This approach does not cover all cases, and it is also prone to errors. I propose the combination of two approaches for handling *ad hoc* compounds.

We can consider the approach such that we have the global implementation for such compounds, the last part of which is listed in two sub-lexicons (adjectives and nouns) in the morphological lexicon. This system allows any noun, adjective, number, numeral, or adverb as first member of the compound.

Its weakness is that it selects just the first gloss of the first member, which is not always correct.

Therefore, we need to list the problematic compounds as such into the lexicon and assign the gloss particularly for that combination. This approach produces correct results, but it is tedious to maintain.

If we have both approaches simultaneously in use, we get competing analyses. Very often the end result of both approaches is the same, but not always. Especially difficult cases produce different results. How can we select the correct interpretation in each case?

The safe solution is to choose that alternative, which was produced by writing the compound directly into the lexicon. Therefore, the alternative described in this report is complementary, and it handles cases, which have no analysis using the normal routines.

Examples of problematic cases are in (21).

(21)
```
"<three-time>"
      "three-time" { kolme N8 } A [ kertainen N38 ]
      "three-time" { kolme N8 } A [ aikainen N38 ]
      "three-time" { kolminkertainen NEN N38 } A
"<long-time>"
      "long-time" { pitkä N10 FRONT } A [ kertainen N38 ]
      "long-time" { pitkä N10 FRONT } A [ aikainen N38 ]
"<20-time>"
      "20-time" A [ kertainen N38 ]
      "20-time" A [ aikainen N38 ]
"<four-time>"
      "four-time" { neljä N10 FRONT , -neljä N10 FRONT COMB , -
neljäs N45 FRONT COMB ORD } NUM-PL A [ kertainen N38 ]
      "four-time" { neljä N10 FRONT , -neljä N10 FRONT COMB , -
neljäs N45 FRONT COMB ORD } NUM-PL A [ aikainen N38 ]
      "four-time" { nelinkertainen NEN N38 } A
```

The English word *time* is ambiguous, and it means *aika* or *kerta* in Finnish. When it is in the compound interpreted as an adjective, it means *aikainen* or *kertainen*. When the word *time* is joined with a number or numeral, it means *kertainen*. Otherwise it means *aikainen*. Using these criteria, we can disambiguate the examples (22).

(22)
```
"<three-time>"
      "three-time" { kolme N8 } A [ kertainen N38 ]
```

```
      "three-time" { kolminkertainen NEN N38 } A
"<long-time>"
      "long-time" { pitkä N10 FRONT } A [ aikainen N38 ]
"<20-time>"
      "20-time" A [ kertainen N38 ]
"<four-time>"
      "four-time" { neljä N10 FRONT } NUM-PL A [ kertainen N38 ]
      "four-time" { nelinkertainen NEN N38 } A
```

Now all examples except for the first and last one have only one reading. In the examples with two readings, both are in fact correct, although they are slightly different. The first reading was produced using the technique that we have been discussing here. The second reading was produced following the following procedure: The compound *three-time* was not recognized by the analyzer, but yet it was interpreted as adjective by the guesser. Then the gloss *kolminkertainen* was added to it in conjunction with the normal gloss adding rules, which were manually prepared and checked.

Now there are two competing interpretations and we must choose one of them. It is clear that the gloss assignment that goes through manual checking is more reliable. Therefore, we select the second interpretation. We also process the readings further (23).

(23)

```
"<three-time>"
      "three-time" { kolminkertainen NEN N38 } A
"<long-time>"
      "long-time" { pitkäaikainen N38 ]
"<20-time>"
      "20-time" A [ 20-kertainen N38 ]
"<four-time>"
      "four-time" { nelinkertainen NEN N38 } A
```

5 Conclusion

For the morphological analyser, *ad hoc* compounds are as any out-of-vocabulary word. However, it is possible to improve greatly the performance of the system by handling the last parts of the compounds as such lexical units, which are allowed to be suffixed to selected POS categories. This does not, however, replace the need to list problematic cases directly to the lexicon, which will then be processed as normal words.